DETERMINING FORM IDENTIFICATION THROUGH THE SPATIAL RELATIONSHIP OF INPUT DATA

Cross-Reference to Related Applications

This application claims the benefit of the filing date of U.S. Patent Application 60/242369, entitled Determining Form Identification Through the Spatial Relationship of Input Data.

Background of the Invention

Field of the Invention:

This invention relates generally to the identification of form documents. In particular, the invention provides a method for identifying form documents through received data.

Description of Related Art:

Automatic forms processing systems utilize computer hardware and software to scan forms, identify the form being processed and extract information from the form. Form identification is an essential step in the processing of forms. It represents the ability of the system to automatically select the correct computer-stored form template, which is used to process the current form. Existing forms processing systems utilize graphic information present on the scanned form to identify the computer-stored form template (form template). Such information may include lines and graphics, blocks of text on the form template, special symbols and form identification numbers.

Some form processing systems receive data apart from the form, although a form may have been used to assist in entering data. With these form-processing systems, only the graphic data entered by the user such as name, address and marked choices is electronically transferred to the forms processing system. Thus, none of the graphic features specific to the design of the processed form are available for form identification. A process described below is utilized to identify the form template, in one embodiment, by relying exclusively on

the locations of fields filled in by the user. [In another embodiment, the probable content type of those fields is used to determine the form template.]

Summary of the Invention

In one aspect of the present invention, there is an automated form identification method, comprising storing electronic forms data in a mobile device, receiving the electronic forms data from the mobile device at a forms processing computer, storing a plurality of form templates on the forms processing computer, each form template having a plurality of entry fields and a layout, the layout identifying a form location for each of the entry fields, and identifying a matching form template for the received electronic forms data based on the entry field locations. One of the form templates may have a first field category and a second field category. The method may additionally comprise scoring each of the fields of the form template depending on the field category and whether data is entered for the field. The scoring may include accumulating field scores in a template score, where the form template with the best score may be declared to be the matching template for the received electronic forms data. The first field category may be indicative of a must fill field and the second field category may be indicative of an optional field. Fields assigned to a first field category or second field category may be linked to other fields such that the outcome of a field assigned to a first field category or second field category indicates the category of the other fields. The method may additionally include scoring each of the fields of the form template depending on the field type and whether data of that type is entered for the field. The method may further include scoring each of the fields having a fixed field length based on the character count of the field.

In another aspect of the present invention, there is an automated form identification system, comprising a mobile device capable of storing electronic forms data, a forms processing computer capable of receiving the electronic forms data from the mobile device, a plurality of form templates stored on the forms processing computer, each form template having a plurality of fields and a unique graphic layout, and a form identification processor operating on the forms processing computer, the form identification processor identifying the best matching form template for the received electronic forms data without use of graphic

elements. One of the form templates may have a first field category and a second field category. Each of the fields of the form template may be scored depending on whether data is entered for the field, where the field scores may be accumulated in a template score. The form template with the best score may be declared to be the best matching template for the received electronic forms data. The first field category may be indicative of a must fill field and the second field category may be indicative of an optional field. The form identification processor may additionally determine the best score by including scoring of the fields of the form template in response to the field type and whether data of that type is entered in the field. The form identification processor may additionally determine the best score by including a scoring of the fields of the form template in response to the character count of the fields for fields having a fixed field length.

In another aspect of the present invention, there is an automated form identification method, comprising receiving electronic forms data from a forms device at a forms processing computer, storing a plurality of form templates on the forms processing computer, each form template having a plurality of entry fields and a layout, the layout identifying a form location for each of the entry fields, and identifying a matching form template for the received electronic forms data based on the entry field locations. The forms device may include a portable digital notepad or a dropout scanner.

In another aspect of the present invention, there is an automated form identification method, comprising receiving electronic forms data at a dropout scanner, transferring the electronic forms data from the dropout scanner to a forms processing computer, storing a plurality of form templates on the forms processing computer, each form template having a plurality of entry fields and a layout, the layout identifying a form location for each of the entry fields, and identifying a matching form template for the electronic forms data based on the entry field locations.

Brief Description of the Drawings

Figure 1 shows the top-level flow of a forms processing method.

Figure 2 is a graphic representation of a forms processing method initiated with a dropout scanner.

Figure 3 is a graphic representation of a forms processing method initiated with a mobile digital signal-receiving device.

Figure 4 displays a sample form used in the forms processing method.

Figure 5 shows the detailed flow of the form identification process.

Detailed Description of the Invention

Figure 1 depicts a forms processing scheme for extracting data from filled forms. The process uses a dropout scanner 102 or mobile digital signal-receiving device 104 to create a bitmap image of the data 106. The bitmap image undergoes image cleaning and processing 108 prior to being submitted to a form identification process 110. The form identification process utilizes a form template having defined entry field locations 112. The bitmap images of processed forms are searched to determine locations of data entry. The locations of data entry are compared to defined entry locations in stored electronic form templates. A best match form template is identified in response to the comparison. Other form identification steps may be incorporated to provide a greater probability in identifying the best match form template. These steps are discussed in detail below. A forms processing computer may be used to perform these steps. Once the form is identified precise entry field locations are known for the processed form. Knowledge of entry field locations allows processing of fields 114, which primarily includes extracting the entered data. A data export 114 sends the extracted data to a database 118. The creation and interpretation of forms is described in U.S. Patent No. 5,555,101, which is hereby incorporated by reference.

Figure 2 is a graphic representation showing a forms processing system using a mobile digital signal-receiving device 202. Various mobile digital signal-receiving devices may be utilized by the system. In one embodiment, a mobile digital signal-receiving device equipped with a special pen capable of digitally recording graphic information is used to

electronically fill forms. In this embodiment, the mobile device (e.g., CROSSPAD, available from A.T. Cross Company) can be carried around separately from the forms processing computer and is able to electronically store a significant amount of entered digital data. The exemplary CROSSPAD functions as a clipboard, allowing users to write thereon. A paper form or form image may be placed atop the CROSSPAD. The form or form image may be used as an overlay for guiding the user to the proper locations for entering the data. The overlay may be positionally affixed to the CROSSPAD as a notepad or singular piece of paper. In another embodiment, a mobile digital signal-receiving device may embed an image of the form within the device. Only the entered data is digitized for use by the forms processing computer. Neither the form nor form image is utilized. The CROSSPAD receives the data as digital signals from the special pen and stores the data and spatial arrangement of the data in memory. The stored data arrangement represents storage of filled forms 202(a). The stored image data (e.g., a bitmap) is loaded into the forms processing computer 204 at a later time.

The exemplary CROSSPAD mobile digital signal-receiving device is essentially a portable digital clipboard that also functions as a digital signal-receiving and storage device. As the user writes to the form with the special pen, the writings are organized into digitized bitmaps, allowing them to be later uploaded to a forms processing computer. The special pen is a digital pen equipped with a small radio frequency (RF) transmitter to allow users to complete the forms affixed to the digital clipboard. An exemplary (unfilled) form that may be one sheet of the notepad tablet or a singular piece of paper is shown in Figure 4. The RF transmitter sends pen stroke data to the clipboard device automatically when the pen makes contact with the form. Signals from the digital pen are stored as digital data (a time series of points) in the memory of the clipboard device. Selected handwriting may be converted to a bit-map for exporting from the device. In one embodiment, the clipboard device may include one MB of flash read-only memory (ROM), which may store up to 50-80 pages of data. The personal computer to which the clipboard device uploads the data uses a Pentium or better processor, and operates using the Windows 95, 98, or NT 4.0 operating software.

In another embodiment, the input device may be what is known as a dropout scanner, such as a model 8125DS scanner available from Bell & Howell. Figure 3 is a graphic representation showing a forms processing system using a dropout scanner 302. The scanner includes a light bulb of a specific color, e.g., red or green. Forms may be printed using a light ink which matches the color of the light bulb. In one embodiment, the entire form is printed using the matching light ink color, which allows for one-color dropout forms. When such a form is completed by a user and then is scanned by the scanner, the scanner does not detect the specific ink color of the form. However, the user enters the data in a different color. The scanner detects and captures the differently colored data. This data is exported as image data (e.g., a bitmap) to a forms processing computer 304.

In both the forms processing systems shown in Figures 2 and 3, the forms processing computer 204/304 receive bitmap images consisting of the data entered through the mobile digital signal-receiving device 202 or dropout scanner 302. This data does not explicitly specify the form template 204(a)/304(a) to be used for processing the form. This data also does not include any of the graphic data utilized for form identification with scanned or faxed forms, such as form identification numbers, lines, preprinted text or other graphic elements from the form template. In one embodiment, image data received by the forms processing system may consist exclusively of the personal information handwritten, hand-typed or electronically entered by the applicant. This information is entered into appropriate fields on the form as specified by the form layout. The form identification process 204(b)/304(b), included in the forms processing computer 204/304, utilizes the location of these entry fields in the form template, compares them with the locations of written text in the currently processed form, and selects the best fitting form template.

In addition to the basic form identification process, other criteria may be included in the process to more accurately identify the correct form. For instance, the data type and data length may be factored into the selection process. Using this additional criterion, form fields are labeled as alpha or numeric. If numeric data entries are found in a field labeled as an alpha field, the identification process is affected accordingly. Another criterion may utilize

fields having an assigned fixed field length. The identification process is weighted accordingly if the number of characters in the field are the same as the assigned field length.

The form identification process is described in greater detail below. Once the form has been identified, interpretation of the fields specific to a particular form and further processing of the data can resume in the forms processing computer 204/304. The extracted data is stored in a database 206/306.

Form Identification Process:

Figure 5 shows a flowchart of the basic form identification process. This reiterative process compares data locations in a data image with corresponding locations of form fields on a form template. A point system establishes a best match for a given data image with respect to two or more form templates. The best match form template is identified as the form and corresponding with the given data input.

More particularly, a first form template 502 having field locating information and field type information is accessed from a form template database 504 having two or more different form templates. The field location identifying information enables identification of the location of the first form field in the bitmap image. Field type information identifies the field classification as either "must fill" or "optional". The field type and location for the first field is identified 506. The location of the first form field in the form template is located 508 in relative correspondence with a datum on a bitmap image 510. A determination 512 is made regarding the existence of a datum within the field boundaries. Depending on the field type, a score is assigned 514(a)/514(b) for the given field location. In one embodiment, if the field type is "must fill", by way of example a score of 1.0 may be given if a datum is found to exist within the field. If there is no datum within the field, a negative score of 0.5 may be assigned. If the field type is "optional", a score of 0.5 may be assigned for an existing datum and 0.0 may be assigned for the nonexistence of a datum. The score is stored in first memory 516 and is added to a preliminary total score in a second memory 518. The process repeats for each field 520 in the form template. The score will increase, decrease or remain the same with each process loop corresponding with each field. When the last field is reached, a total

score is assigned which corresponds to the given form template 522 used for that first template loop. A determination is made if the form template was the last template 524 in the form template database. If it is not, a subsequent form template is accessed and the process is repeated. When the last form template is accessed and analyzed, the form template having the corresponding best score is identified 526 as the form used or corresponding with the given input data. The identified form template 528 is then used in the forms processing system.

The form identification process may be refined further by adjusting the scores of individual fields based on specific characteristics of the fields. In one such refinement, a data type assignment may be given in the form template for a given field. The score for that field may be adjusted according to the type of data recognized at that field location. For example: If a numeric field type is assigned for a data field used for entering a social security number, the score for the form template would be improved if numerals were recognized in that field and weakened if alpha characters were recognized.

Additional scoring weight may be given for finding the correct number of characters or appropriate template characters, such as "/" or "–" in the case of date fields, when the user is allowed to enter such characters.

Different fields on a form template fall into one of the two categories: MUST FILL and OPTIONAL. The MUST FILL fields may be represented by the name of the applicant, address, social security number and so forth. Choice fields, additional names, addresses and numbers may be examples of OPTIONAL fields. On a filled out form, either type of field may be filled or missing (not filled). Referring to Table 1, the following scheme of weights may be adopted to count filled and missing fields, and to compute the score for each form field and total score for each form template in the currently processed set.

Table 1

Field type	Field is filled	Field is missing
MUST FILL	1.0	- 0.5
OPTIONAL	0.5	0.0

For MUST FILL fields, the presence of text at the expected location yields a weight of one, while a missing MUST FILL field contributes a negative number. OPTIONAL fields contribute zero when missed and a number between zero and one when filled. The total score for a form is the sum of weights for all fields in the template. The template with the best score yields the correct form.

To provide a greater degree of accuracy in the form selection process, the scoring can be further weighted according to field characteristics, as described above. Identification of forms using field characteristics is described in U.S. patent application 09/656,719, entitled "Method and System for Searching Form Features for Form Identification", which is hereby incorporated by reference. By way of example, as shown in Table II below, fields may be assigned a specific field type or given a fixed field length. Such fields allow the method to provide the greater degree of accuracy. Referring to Table II, when all the characters of a given field are recognized to be those of the assigned field type, a positive score is added. When a character is found that does not correspond to the assigned field type, the score is decreased a fractional negative number. For fields that are assigned a fixed field length, a positive score is added when the number of characters corresponds to the assigned field length. A negative fractional score is added for each character count difference.

Table II

Field Characteristic	All Correct	Incorrect
ALPHA/NUMERIC	0.5	- 0.2/incorrect char
LENGTH (fixed length fields)	0.5	- 0.2/char count diff